Boffin Access

Open Access Journal | Volume: 1.1

Methods of Microbiology and Molecular Biology

ISSN 2517-7435

Open Reading Frame Filtering Methods for Identification of Genic Sequences

Nayra M. Al-Thani[1]
Simeon S. Andrews[1]
Dietrich Büsselberg[1]*

[1]Weill Cornell Medicine in Qatar, Qatar Foundation-Education City, Doha, Qatar

Abstract

Protein-Protein Interactions (PPIs) help understanding disease processes and their mechanisms. Ideally, researcher would like to understand the full network of PPIs that take place within a cell "interactome", that is why Open Reading Frame (ORF) filtering method was utilized. ORF is a DNA fragment that lacks stop codon and has the potential to demonstrate a physiological interaction. ORFs are obtained by random shearing of DNA into small fragments. Fragments are filtered by insertion upstream of a selectable marker, to allow the survival of cells that only have an ORF. As vast majority of out of frame fragments encodes a premature stop codon. Here we present a method for filtering of genic ORFs 'real gene' which results in a physiological protein. Furthermore, researchers thought fragments libraries rather than full-length libraries are perhaps counterintuitive, as expected to result in a high rate of false negatives. However, researchers have found fragments libraries rather than full-length reduce number of false negative interactions. Lastly, great advantage to note is while using fragmented library it allows localization of interaction site, offering a robust path for drug targets, treatments towards cancer and the arising resistance to antibiotics.

Keywords

Open Reading Frame; ORF filtering; Two-hybrid system; Gene fragment; Protein-protein interaction

Protein-protein Interactions

Most cellular processes are influenced or directly mediated by protein-protein interactions (PPIs). Studying PPIs is therefore essential for understanding normal and pathological physiology within a cell. Understanding PPIs helps us to understand disease processes such as cancer and their mechanisms. Ideally, we would like to understand and study the entire network of protein interactions, which is referred to as the "interactome". An interactome defines the full network of PPIs that take place within a cell [1].

Diverse methods have been used to identify interactomes, including proteomics methods [2]. However, the chemistry and technology of protein quantitation have substantial challenges. In contrast, DNA-sequencing technologies have long been robust, and particularly in the past decade "next-generation" sequencing has revolutionized our ability to quickly and accurately sequence vast amounts of genetic material. Various methods have used DNA readouts to study PPIs. Of these, the two-hybrid system is probably most frequently used [3]. The original system used yeast proteins encoding a DNA-binding domain and an activation domain. One protein of interest ("X") is N-terminally fused to the binding domain, while another protein is similarly attached to the activation domain ("Y"). The binding domain protein binds to a specific DNA sequence and the activation domain recruits transcription factors, which are necessary to initiate the transcription of a reporter gene. Only if the "X" and the "Y" protein interact, bringing the binding domain and the activation domain into proximity, will the reporter gene be transcribed (Figure 1).

The two-hybrid system has been used in the intervening decades, and it has two major variations, which are the "yeast" and the "bacterial" two-hybrid systems.

In 1997 Fromont-Racine and co-workers used yeast 2-hybrid system (Y2H) to identify protein interactions in yeast [4]. Thereafter, the Y2H model was used to screen for PPIs in large scale for species such as Saccharomyces cerevisiae, Helicobacter pylori, Drosophila melanogaster, Caenorhabditis elegans, and Homo Sapiens [5-11]. Three years after the first screening using the Y2H system Joung and colleagues were the first utilizing a bacterial 2-hybrid (B2H) system with a large library (~ 108 in size) [12]. The B2H system has two major advantages compared to the Y2H system, as it has a faster growth rate, and higher transformation efficiency [13].

Gene Random Fragmentation Libraries Versus Full-Length Libraries

Initial experiments with the two-hybrid systems generally employed full-length genes.

Article Information

DOI: 10.31021/mmmb.20181105
Article Type: Review Article
Journal Type: Open Access
Volume: 1 **Issue:** 1
Manuscript ID: MMMB-1-105
Publisher: Boffin Access Limited

Received Date: January 09, 2018
Accepted Date: January 26, 2018
Published Date: January 31, 2018

Citation: Al-Thani NM, Andrews SS, Büsselberg D. Open Reading Frame Filtering Methods for Identification of Genic Sequences. Methods Microbiol Mol Biol. 2018 Jan; 1(1): 105

***Corresponding author:**
Dietrich Büsselberg
Weill Cornell Medicine in Qatar
Qatar Foundation-Education City
POB 24144, Doha, Qatar
Tel no: +974 33480728
E-mail: dib2015@qatar-med.cornell.edu

Copyright: © 2018 Büsselberg D, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 international License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

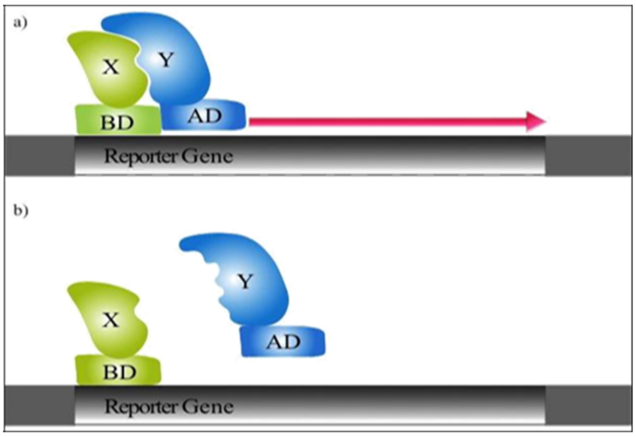

Figure 1: a) Based on the yeast 2-hybrid system (from Fields et al., 1989. 1) in case protein X and Y do interact, their interactions bring together the binding domain (BD) and the activation domain (AD) into contact and drive the expression of the reporter gene. b) In case protein X and Y do not interact, the reporter gene is not transcribed. BD, binding domain that binds to binding site within DNA sequence; AD, activation domain which binds to promoter that drives the expression of reporter gene; X, protein of interest bait; Y, protein of interest prey.

The use of fragments rather than full-length libraries is perhaps counterintuitive, as we know that many protein structures and interactions are impossible with fragments, and might be expected to result in a high rate of false negatives. Surprisingly, researchers have found that gene fragment libraries reduce false negatives interactions rather than full-length gene; and thorough screening reduce false positives interaction [7,14]. When Boxem, M et al, used fragments rather than full-length genes, they found that they recovered more physiological interactions, and in their limited set no false positive interactions [14]. Presumably, this is because fragments of the protein may avoid problems of folding or translocation found by full-length proteins. Any false positive interactions can also be eliminated with more thorough library screening [7].

The use of fragments has the added benefit of permitting more rapid screening, as one does not have to first devise a library of all protein-coding genes with specific primers before testing pairs. Random fragmentation quickly allows cDNA to be converted into testable fragments. Finally, note that the use of fragments allows localization of interactions to specific regions of proteins. Rather than knowing only those two proteins interact, we can define their interacting regions as well. With overlapping fragments, we can even identify the minimal interacting region.

Open Reading Frame Concept

For a fragment to potentially demonstrate a physiological interaction, it must be an open reading frame (ORF) (Figure 2). An ORF is a DNA sequence without a stop codon and has the potential to encode proteins. If DNA is sheared into random fragments, followed by insertion into a vector, the majority of gene fragments (83%) do not represent any functional gene (termed as out of frame), therefore coding non-physiological proteins. Out of the 6 possible frames (Figure 3) only 1 fragment corresponds to the gene frame, which encodes the physiologically relevant protein. Therefore, these fragments need to be filtered in order to discard those that are non-physiological. This process of removing non-ORF sequences is what is termed "ORF filtering."

All methods currently in use for ORF filtering rely on the underrepresentation of stop codons in truly coding sequences. A truly random DNA sequence will, on average, encode a stop codon every 21 triplets. Indeed, with only 63 codons, there is a 95% probability of having at least one stop codon if the sequence is random, and with fragments encoding 100 amino acids (300 bp), there is a 99 % chance that random sequence will have a stop codon. By contrast, a coding sequence of DNA will of course avoid stop codons until the end of the sequence has been reached. If we take fragments of cDNA just 300 bp long, but in random frames, then 5/6 (83.3%) will be in the wrong frame. Yet 99% of those will have a stop codon; if we can selectively eliminate fragments with stop codons, the in-frame percentage of the library will go from 16.7% up to 96%. This has the disadvantage of selecting against in-frame sequences that include the physiological stop codon, and thus the C-termini of proteins are expected to be under represented.

In order to filter and express ORFs, a sheared DNA fragment is cloned upstream of a selectable marker. If it is an ORF, then the selectable marker will be transcribed and expressed in those cells. The vast majority of DNA fragments that are out of frame will encode a premature stop codon, and consequently the selectable marker is

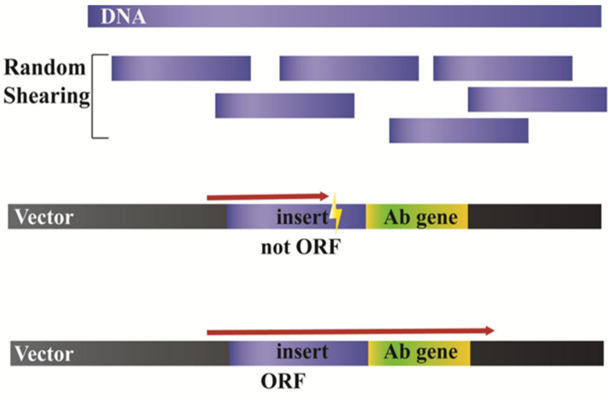

Figure 2: Random shearing of DNA and insertion into vector, can yield gene expression of antibiotic resistance gene (Ab gene) only in case the insert lack stop codon (which means it's an Open Reading Frame ''ORF''). This method showed high filtration for inserts that are in-frame 1 compared to the five other frames.

```
a) 1 ATG ATT GAA CAA GAT GGA TTG CAC GCA GGT TCT CCG GCC GCT TGG GTG GAG AGG CTA TTC
61 GGC TAT GAC TGG GCA CAA CAG ACA ATC GGC TGC TCT GAT GCC GCC GTG TTC CGG CTG TCA
121 GCG CAG GGG CGC CCG GTT CTT TTT GTC AAG ACC GAC CTG TCC GGT GCC CTG AAT GAA CTG
181 CAA GAC GAG GCA GCG CGG CTA TCG TGG CTG GCC ACG ACG GGC GTT CCT TGC GCA GCT GTG
241 CTC GAC GTT GTC ACT GAA GCG GGA AGG GAC TGG CTG CTA TTG GGC GAA GTG CCG GGG CAG
301 GAT CTC CTG TCA TCT CAC CTT GCT CCT GCC GAG AAA GTA TCC ATC ATG GCT GAT GCA ATG
361 CGG CGG CTG CAT ACG CTT GAT CCG GCT ACC TGC CCA TTC GAC CAC CAA GCG AAA CAT CGC
421 ATC GAG CGA GCA CGT ACT CGG ATG GAA GCC GGT CTT GTC GAT CAG GAT GAT CTG GAC GAA
481 GAG CAT CAG GGG CTC GCG CCA GCC GAA CTG TTC GCC AGG CTC AAG GCG AGC ATG CCC GAC
541 GGC GAG GAT CTC GTC GTG ACC CAT GGC GAT GCC TGC TTG CCG AAT ATC ATG GTG GAA AAT
601 GGC CGC TTT TCT GGA TTC ATC GAC TGT GGC CGG CTG GGT GTG GCG GAC CGC TAT CAG GAC
661 ATA GCG TTG GCT ACC CGT GT ATT GCT GAA GAG CTT GGC GGC GAA TGG GCT GAC CGC TTC
721 CTC GTG CTT TAC GGT ATC GCC GCT CCC GAT TCG CAG CGC ATC GCC TTC TAT CGC CTT CTT
781 GAC GAG TTC TTC TGA
```

```
b)
Frame1      GCT ATG     ATT GAA............TTC TTC TGA
Frame 2     TGC TAT     GAT TGA......ATT CTT CTG A
Frame 3     CTG CTA     TGA TTG...AAT TCT TCT GA
Frame- 1    GCT AGT     CTT CTT............AAG TTA GTA
Frame- 2    TGC TAG     TCT TCT......TAA GTT AGT A
Frame- 3    CTG CTA     GTC TTC...TTA AGT TAG TA
```

Figure 3: a) Shows example of Kanamycin gene sequence in the form of triplet (Using: www.bioinformatics.org/sms2/group_dna.html) labeled red ATG the start codon and labeled blue TGA stop codon. b) Shows how the frames might be changed due insertion of random fragments of genes. If it's frame 1 the sequence will be expressing physiologically relevant protein. While if it's frame 2 the expression will be shifted by 1 base pair, leading to expressing non-physiological protein. Similar issue will be encountered for frame 3 (shifted by 2 base pair) and those three frames could be inserted in the wrong orientation leading to frames -1, -2, -3. Where the gene is read from opposite direction (i.e. from the end to start).

not translated, resulting in a difference in selectability between ORF-containing cells and those without ORF fragments. For instance, if an antibiotic resistance gene is the selectable marker, then only the cells with ORF fragments will survive in the presence of antibiotic (Figure 4) [15]. This method was adopted by Weinstock, G. M., and co-workers (1983), who inserted random fragments into a vector between the outer membrane protein (Omp) gene and beta-galactosidase (LacZ) gene (which can be used to screen via blue-white colony selection). The vector has LacZ(-), which corresponds to the nonfunctional gene [16]. However, by inserting of random-fragments that realign both genes Omp and LacZ, the LacZ(+) becomes functional and is expressed on the colonies and, therefore, can be selected by blue-white screening. Furthermore, the libraries with random-fragments were enriched from 54% to 100% ORFs by the selection of an antibiotic [17,18]. Therefore, such a selection improves the quality of the library [19]. Moreover, this method is also capable of localizing the sites of interactions within the sequence [20,21].

There are several organisms that can be used for ORF filtering including: bacterial strains, viruses such as phages, yeast. All these methods are based on a series of experiments, starting by 1) isolating a gene, 2) shearing to ORF fragments, 3) amplifying by polymerase chain reaction (PCR), 4) ligation into a vector, 5) transfection into cell, 6) applying selective pressure to the library, and 7) sequencing the targeted DNA.

ORF Selection in Bacteria (*Escherichia coli*)

To identify ORF's in bacteria a marker, such as AmpR, (marker for ampicillin resistance), is used to test the presence of ORFs. DNA fragments are inserted upstream of AmpR gene and downstream of its leader sequence. The leader sequence allows the export of the transcribed product of AmpR gene to the periplasm (Figure. 4), its site of action. Different antibiotics (e.g. chloramphenicol, kanamycin, spectinomycin, tetracycline) can be used as selectable markers [22]. Furthermore, some methods use a more complex cloning by insertion of some sequences, such as Lox sequence, which is cleaved by the Cre recombinase [18,23]. This allows a recombination of the ORF and the formation of the fused DNA product with a tag gene. By flanking the ORFs with recombination elements, we can facilitate the isolation of ORFs for further studies and validation of ORF interactions [24,25].

The first time the frame concept was used to generate a MH3000 E.coli strain which had a -galactosidase (LacZ) gene out-of-frame [16]. These researchers inserted ORFs downstream the OmpF gene and upstream of an out-of-frame LacZ gene. When the fragments were inserted, those which changed the frame to generate a functional LacZ gene could produce blue colonies in the presence of X-Gal. These blue colonies could be verified to contain functional proteins.

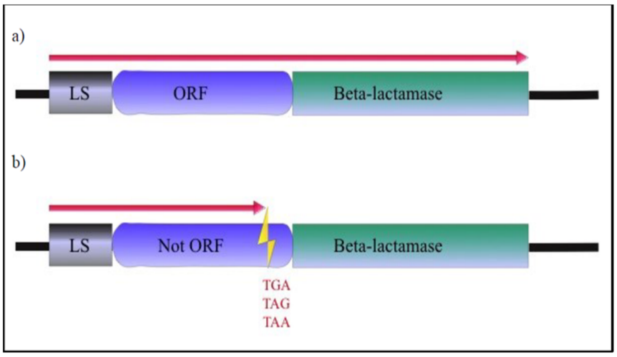

Figure 4: Method of filtering ORFs using beta-lactamase gene, which is ampicillin resistance (Di Niro et al,. 2010; D'Angelo et al,. 2011). ORF is representing fragment insert that are tested for lacking stop codon to be filtered. LS; stands for leader sequence were the transcription of the sequence starts. a) Represents that transcription of the whole sequence till the ampicillin resistance gene (beta-lactamase), since the cDNA lack stop codon (ORFs). b) Only a part will transcribe as a stop codon was introduced. Colonies that confer such a stop codon will not survive in ampicillin, since they are not producing the ampicillin resistance gene.

Davis & Benzer showed that ORF frequencies are dependent on the concentration of antibiotic concerning their selection, library size, or bacterial strain [26]. They showed 8% of the clones are in frame before selection, while this fraction increased to 70% following the selection. Furthermore, selection frequency differed from ORFs library size. Both strains XL1-Blue and DH10B are capable of cloning larger fragment, however XL1-Blue resulted in higher transformation efficiency compared to DH10B. They concluded that for a smaller library size a higher concentration of the antibiotic did result in a better selection, while this was opposite for large library seizes. By chance the ORF fragment could be orientated in the wrong orientation. To overcome this issue Davis and co-workers used directional cloning using two different restriction enzymes to clone the ORFs into the expression vector [26]. Moreover, they used PCR primers to modify kanamycin gene by having a stop codon in the second reading frame of ATGA. This allowed them to ensure the ORFs will not survive if the reading frame starts from the second nucleotide.

Four test genes were used to confirm the "theory of frame" by shifting two of the genes, adding one or two bases [22]. The vector used had a Chloramphenicol resistance, and an ampicillin resistance gene. The four test genes were inserted upstream of AmpR gene and were transformed and plated in two different plates (Chlor and Amp). The colonies with frame-shifted genes do not survive in Ampicillin plates since it does not have a functional AmpR gene. Therefore, all the four test gene were able to grow in Chloramphenicol plates.

Filtering of genic ORFs for a 'real gene' – resulting in a physiological protein - researchers used a vector that has a chloramphenicol resistance to grow colonies on plates [27]. Thereafter, they harvested cells and grew them in selective media supplemented with both chloramphenicol and with different concentrations of ampicillin (as a selective marker). This step was followed by sequencing to identify ORFs which obtained 96% corresponding to real genes. Statistical analysis showed that the activity of beta-lactamase rises with increasing concentrations of ampicillin. This proves that higher expression of beta-lactamase is essential for colonies to grow in high concentration of ampicillin.

ORF Selection in Viruses (T7 Phage)

The phage display method inserts the gene (which encodes the protein of interest) into a phage coat protein that is expressed on the surface of the phage. Thereafter the gene is expressed in bacteria (as a host); a process called transduction. The primary bacteriophages used for phage display are T7 and M13, both of which can use Escherichia coli as a host.

In one of the first reports of ORF filtering for phage display, researchers modified a vector by eliminating the original multiple cloning site (MCS) inserting a new site through inverse PCR [19]. This step allowed them to design ORFs which are in-frame even when sub-cloned into derivatives vectors, which they constructed. After ligation of the fragments upstream of AmpR, the vector was transformed into an XL-1blue strain and ampicillin selection was applied. In order to display the ORF's on the phage surface, the insert was cloned into a derivative vector and was transformed into a bacterial strain (ER2738) to grow under kanamycin selection. Phagemid were rescued by helper phage and sequencing analysis of those samples determined that 97% were ORFs.

Zacchi et al fused fragment upstream of the beta-lactamase to filter ORFs and flanked the insert by lox recombination sequence [18]. Following ampicillin selection, they excised the beta lactamase gene from the vector. To facilitate the purification of those ORFs using Phage display, the constructed vector had fd phage tag "gene 3". Results confirmed using high concentration of antibiotic eliminate out-of frame sequences, (when using 12 μM ampicillin they had 100% ORFs and 0.2% out of frame; but when using 25 μM ampicillin they had 85% ORF and none out of frame). From the 100% filtered library, 80% was detected using Dot blot for protein detection and the mapped ORFs represented 50% genic ORFs.

In 2010, Di Niro et al, applied the antibiotic screening technique to prepare ORFs for expression in phage display vectors in a more seamless way. Genes were fragmented into to 100-600bp, and fragments were cloned in the right orientation using restriction enzymes sites. The ORFs fragments were inserted upstream of AmpR gene, but with lox sequences flanking the AmpR sequence. Downstream from the AmpR and Lox sequence is a g3p gene, which encodes a phage coat protein used to display ORFs in phage surface. Once ligated, the vectors were transformed into bacteria and grown on ampicillin-containing plates for ORF selection. Positive clones were then transformed into a bacterial strain that has a constitutively active Cre-recombinase. This cleaves the lox sequence and recombines ORFs with g3p gene, eliminating the AmpR gene. In this way, the ORF only needs to be once cloned into the vector, while still permitting expression of the ORF/g3p fusion without AmpR. To validate interactions ORFs were displayed in phage for the enzyme transglutaminase 2 (TG 2). This method allowed the selection of 99% ORFs in which 85% corresponding the correct frame of the gene, and provided the local regions of interactions domain.

By contrast, Caberoy NB et al, used phage display itself to select ORFs [28]. They used the T7 phage, inserting their ORF at the C-terminal end of the Capsid 10B protein. Crucially, however, they added a 3C protease cleavage site and then a biotinylation site further downstream. Consequently, only virus particles encoding an ORF will be biotinylated. The cDNA library was selected using streptavidin to isolate biotinylated ORFs, and these were cleaved from resin with the 3C protease. The recovered library was re-amplified in bacteria to generate an ORF-selected library suitable for use in selection experiments. They found 17 ORFs, of which 13 encode different protein were selected using phage display. Phage display of cDNA library fused with biotinylation tag in the C-terminus confirmed following selection that clones had 90% enriched ORFs inserts [29].

Gene sheared, gel purified fragments of 100-300bp. Preformed ampicillin selection to filter ORFs, following step vector transformed into strain that express constitutively active Cre gene to remove ampicillin gene from ORFs after selection [30]. Phage display of ORFs by infecting the bacteria with helper phage M13K07 represented 94% ORFs library.

Fragmented gene into 200-800bp via sonication, those fragments was cloned into vector [31]. Followed by transform into strain resistance to chloramphenicol and ampicillin as selective marker. To confirm that the target sequences obtained, the samples were sequenced and to determine the structure of the enzyme crystallization performed. For the purification and crystallization of proteins, a His-tag was attached in the N-terminus to have protein in the soluble form. Following this approach, they were able to identify two domains on the gene, covered 739 genes from chromosome 1 and 540 genes from chromosome 2 with a total of 1279 ORFs in their library.

Open reading frame percentage (ORFs%) corresponds to percentage of sequences that were isolated without having a stop codon. (ORFs genic%) is percentage of ORF sequences that were isolated without having a stop codon and when aligned to the reference gene it aligns to the correct gene frame. The (selective marker) is the marker that have been used in the ORF filtering vector to select for ORF sequence and filter out the one with stop codon based on antibiotic selective pressure for bacteria and yeast, Tag presence for phage. Fragments size used for sheared DNA library on base pair (bp), Application applied such as Phage display to and further tests applied to validate ORF (Validation methods). Lastly, the strain used for selection whether it has been done in bacteria, phage, or yeast and corresponding authors name. Authors highlighted yellow are for ORF selection using bacteria, while green using yeast, and blue using phage.

ORF Selection in Yeast (*Saccharomyces Cerevisiae*)

ORFs have also been tested in yeast, by transforming first into a bacterial strain for expression of the desired vector, which must be ampicillin resistance by plating into Amp plates [32]. Following the selection of the desired vector, the plasmid was transformed into yeast to get ORFs through histidine induction medium to select ORFs that are tagged with histidine gene. The ORFs were tagged with histidine gene, to filter out the ORFs. As other researched used ampicillin as selective marker, here Holz and his colleague used histidine gene as selective marker for yeast. Through this experiment they were able to cover 60% ORFs.

Organism Selection

Using bacteria, as a host to obtain ORFs is more efficient and reliable compared to phage and yeast. In order to get an ORF library using Phage display it requires enrichment via multiple rounds of purification and amplification [29]. The only advantage of that method as it allows purification of larger fragments [13,30]. Although Caberoy and her colleagues were able to achieve 90% ORFs it does not correspond to real genic ORF%, since ORF is just a sequence that lacks stop codon. However, when some of ORF sequences aligned to the corresponding gene, it does not align to the correct frame of the gene. More importantly, ORF filtered library using bacteria provides faster growth rate and higher transformation efficiency [12]. Yeast ORF filtered library is not yet extensively tested as bacteria, Holz C et al, applied ORF filtering on yeast were they achieved 60% ORFs, using insert of 200-20000bp [32,33]. However, given that yeast generally have much lower transformation efficiencies, they are unlikely to be an ideal host except under specific conditions, such as if a fragment is not believed to fold properly except in a eukaryotic environment.

Bacterial Strain

From Table 1 most of the ORF selection using bacteria is done using DH5alphaF' strain, since it has a high plasmid yield and high transformation efficiency (which is 1×109cfu/μg). Furthermore, (recA gene) responsible for heterologous recombination is mutant, ensuring high stability of the insert. Plus DH5alphaF' lacks some endonucleases that will start digesting the plasmid during the isolation process [33].

Ampicillin Antibiotic as a Selectable Marker

A good note from Table 1 is that ampicillin is the most common antibiotic used as selective marker for ORF filtering. One reason for this is that it requires the expression of the antibiotic resistance gene with its leader sequence. If the fragment is inserted between them, this ensures that the resistance expression is from the full inserted sequence of the ORF. While the other antibiotics, which lack a leader sequence (since they are not exported), will allow the expression of some out-of-frame fragments which may have alternate translation start sites. Thus, the leader sequence of ampicillin ensures that all of the expressed ORFs contain the full insert sequence.

Fragment Library Insert Size

A higher ORF percentage is achieved using smaller insert fragments. Using bacteria ORF filtered library achieved highest ORF percent when utilized insert size ranging 100-800bp (Table1). However, a fragment ranging from 100-500bp is better in the sense that a fragment with size of 300bp will have 99% chance of having a stop codon. The great advantage of using fragmented library is that it allows localization of interaction site [20,21]. The ORF filtering eliminates fragments with stop codon, providing good selection for more genic ORFs (Figure 5). When fragmented libraries were used with ORF filtering a recovery of more physiological interactions was achieved [14].

ORF filtering is a great tool for providing functional sequence within the gene. This offers a robust path for discovery of drug targets, treatment of infections especially resistance to antibiotic and cancer.

Acknowledgements

The authors thank Dr. Joel Malek (Genomics core at WCM-Q) for sharing his extensive knowledge in the field of genomics research.

	Genic %	Selective marker (µg/ml)	Fragment	Application	Validation methods	Selection	
total of 1279 ORFs	-	Amp 100	200-800	Phage display	PAGE, Crystallization	BL21(DE3	Gourlay, et al [31]
76	96	Amp ranging 0.25 to 100	200-800	-	Functional Assay β-lactamase	DH5αF'	D'Angelo, et al [27]
99	85	Amp 15	100-600	Phage display	ELISA, PCA	DH5αF'	Di Niro, et al [23]
94	88	Amp 12	100-300	Phage display	ELISA, Western Blot	DH5alphaF	Di Niro, et al [30]
100	50	Amp 12	100-300	Phage display	ELISA, Dot Blot	DH5αF', BS1365	Zacchi et al [18]
100	-	Amp 100	600-650	-	-	DH5αE'	Lutz, et al [22]
60-80	-	Kan ranging 6.25 to 100	100-300	-	Anti-GST Western Blot	DH10B, XL1-blue	Davis and Benzer [26]
82-96	-	Amp 25	400-500	-	Protein Blot, β-galactosidase Assay	LG90	Gray, et al (1982)
60	-	Histidine 40mg/ml	200-2000	-	Protein Expression	GRF18	Holz, et al [32]
90	-	Biotin	300-1500	Phage display	ELISA	T7 Phage	Caberoy, et al [29]
90	-			Phage display	ELISA, Western Blot	T7 Phage	Caberoy, et a

Table 1: Summary of ORF filtering methods from literature

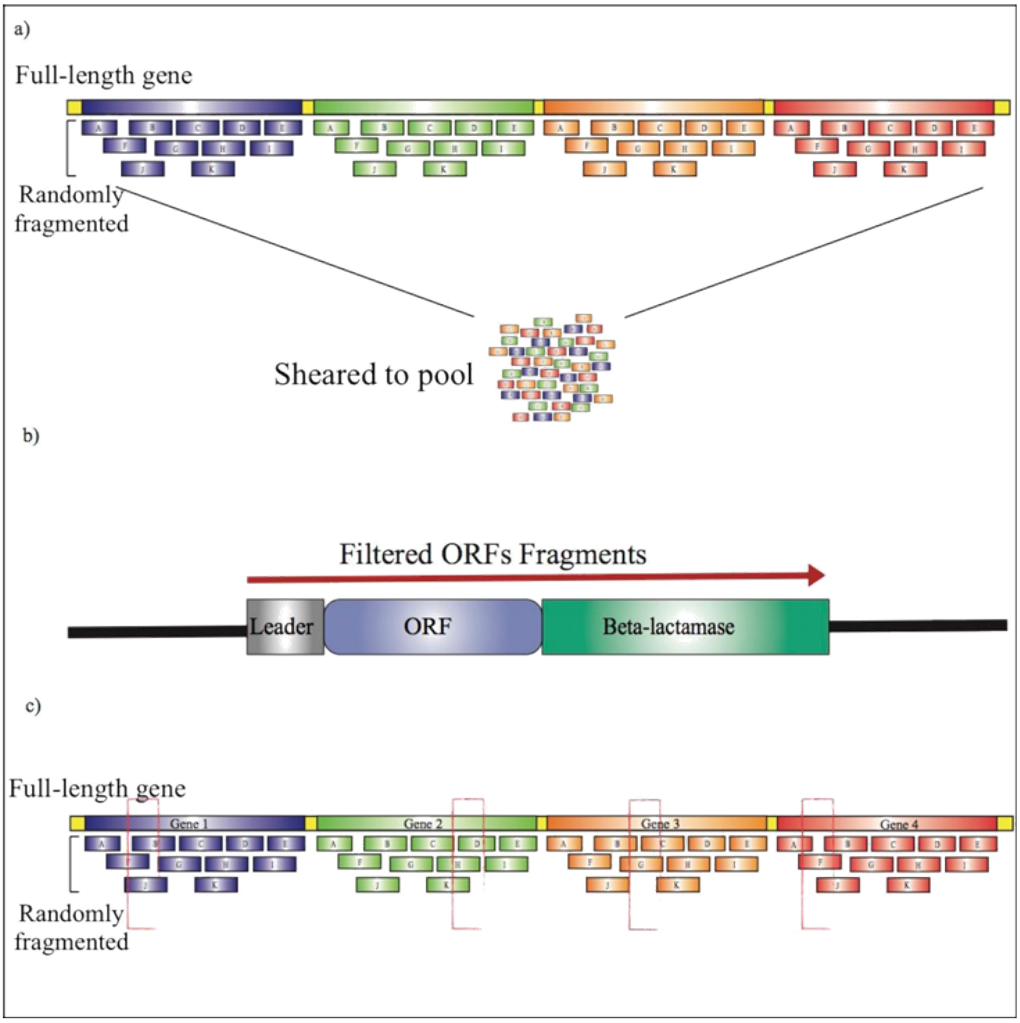

Figure 5: a) DNA sequence presenting of four genes (full-length) that are randomly fragmented and sheared to pool of DNA fragments. The pool is representing a whole mixture of all possible sequences fragments within the genome. b) The pooled fragments are inserted into ORF expression vector to filter out only the fragments that does not have a stop codon (which allow the survival under antibiotic selection). c) Finally when ORF filtered library is validated for the interaction and aligned to the reading frame using bioinformatics tools it allow the localization of the interaction site within the gene. As showing for gene.1 the interaction site is covered from B, F, and J. whereas for gene.2 is covered by fragment D, H, and K. gene.3 fragment B, C, G, and a partial part of K which could be presented as weak interaction during the test. The same for gene.4 when interaction is applied, the strongest interaction is going to be from F fragment, then J while A and B will be weaker.

Conflict of Interest

The authors declare no competing interest

References

1. Cusick ME, Klitgord N, Vidal M, Hill DE. Interactome: gateway into systems biology. Hum Mol Genet. 2005 Oct; 14 Spec No. 2:R171-181.
2. Snider J, Kotlyar M, Saraon P, Yao Z, Jurisica I, et al. Fundamentals of protein interaction network mapping. Mol Syst Biol. 2015 Dec;11(12):848
3. Fields S, Song O. A novel genetic system to detect protein-protein interactions. Nature. 1989 Jul;340(6230):245-246.
4. Fromont-Racine M1, Rain JC, Legrain P. Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. Nat Genet. 1997 Jul;16(3):277-282.
5. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, et al. 2000. A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae. Nature. 2000 Feb;403(6770):623-627.
6. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, et al. 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci U S A. 2001 Apr;98(8):4569-4574
7. Rain JC, Selig L, De Reuse H, Battaglia V, Reverdy C, et al. 2001. The protein–protein interaction map of Helicobacter pylori. Nature. 2001 Jan;409(6817):211-215.
8. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, et al. A protein interaction map of Drosophila melanogaster. Science. 2003 Dec;302(5651):1727-1736
9. Li S, Armstrong CM, Bertin N, Ge H, Milstein S, et al. A map of the interactome network of the metazoan C. elegans. Science. 2004 Jan;303(5657):540-543.
10. Colland F, Jacq X, Trouplin V, Mougin C, Groizeleau C, et al. Functional proteomics mapping of a human signaling pathway. Genome Res. 2004 Jul;14(7):1324-1332
11. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, et al. Towards a proteome-scale map of the human protein-protein interaction network. Nature. 2005 Oct;437(7062):1173-1178.
12. Joung JK, Ramm EI, Pabo CO. A bacterial two-hybrid selection system for studying protein-DNA and protein-protein interactions. Proceedings of the National Academy of Sciences of the United States of America. 2000 Jun;97(13):7382–7387.
13. Wolfe SA, Greisman HA, Ramm EI, Pabo CO. Analysis of zinc fingers optimized via phage display: evaluating the utility of a recognition code. J Mol Biol. 1999 Feb;285(5):1917-1934.
14. Boxem M, Maliga Z, Klitgord N, Li N, Lemmens I, et al. 2008. A protein domain-based interactome network for C. elegans early embryogenesis. Cell. 2008 Aug;134(3):534-545.
15. Seehaus T, Breitling F, Dübel S, Klewinghaus I, Little M. A vector for the removal of deletion mutants from antibody libraries. Gene. 1992 May;114(2):235-237.
16. Weinstock GM, ap Rhys C, Berman ML, Hampar B, Jackson D, et al. Open reading frame expression vectors: A general method for antigen production in Escherichia coli using protein fusions to beta-galactosidase. Proceedings of the National Academy of Sciences. Proc Natl Acad Sci U S A. 1983 Jul;80(14):4432-4436.
17. Rombel IT, Sykes KF, Rayner S, Johnston SA. ORF-FINDER: A vector for high-throughput gene identification. Gene. 2002 Jan;282(1-2):33-41.
18. Zacchi P, Sblattero D, Florian F, Marzari R, Bradbury AR. Selecting open reading frames from DNA. Genome Res. 2003 May;13(5):980-990.
19. Faix PH, Burg MA, Gonzales M, Ravey EP, Baird A, et al. Phage display of cDNA libraries: Enrichment of cDNA expression using open reading frame selection. Biotechniques. 2004 Jun;36(6):1018-1022.
20. Reboul J, Vaglio P, Rual JF, Lamesch P, Martinez M, et al C. elegans ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. Nat Genet. 2003 May;34(1):35-41.
21. Malek JA, Wierzbowski JM, Tao W, Bosak SA, Saranga DJ, et al. Protein interaction mapping on a functional shotgun sequence of Rickettsia sibirica. Nucleic Acids Res. 2004 Feb;32(3):1059-1064.
22. Lutz S, Fast W, Benkovic SJ. A universal, vector-based system for nucleic acid reading-frame selection. Protein Eng. 2002 Dec;15(12):1025-1030.
23. Di Niro R, Sulic AM, Mignone F, D'Angelo S, Bordoni R, et al. Rapid interactome profiling by massive sequencing. Nucleic Acids Res. 2010 May;38(9):e110.
24. Hartley JL, Temple GF, Brasch MA. DNA Cloning Using In Vitro Site-Specific Recombination. Genome Res. 2000 Nov;10(11):1788-1795.
25. Siegel RW, Jain R, Bradbury A. Using an in vivo phagemid system to identify non-compatible loxP sequences. FEBS Lett. 2001 Sep;505(3):467-473.
26. Davis C, Benzer S. Generation of cDNA expression libraries enriched for in-frame sequences. Proceedings of the National Academy of Sciences of the United States of America. Proc Natl Acad Sci U S A. 1997 Mar;94(6):2128–2132.
27. D'Angelo S, Velappan N, Mignone F, Santoro C, Sblattero D, et al. Filtering "genic" open reading frames from genomic DNA samples for advanced annotation. BMC Genomics. 2011 Jun;12 Suppl 1:S5.
28. Caberoy NB, Zhou Y, Alvarado G, Fan X, Li W. Efficient identification of phosphatidylserine-binding proteins by ORF phage display. Biochem Biophys Res Commun. 2009 Aug;386(1):197-201.
29. Caberoy NB, Alvarado G, Li W. Identification of calpain substrates by ORF phage display. Molecules (Basel, Switzerland). 2011 Feb;16(2):1739-1748.
30. Di Niro R, Ferrara F, Not T, Bradbury AR, Chirdo F, et al. Characterizing monoclonal antibody epitopes by filtered gene fragment phage display. Biochem J. 2005 Jun;388(Pt 3):889-894.
31. Gourlay LJ, Peano C, Deantonio C, Perletti L, Pietrelli A, et al. Selecting soluble/foldable protein domains through single-gene or genomic ORF filtering: structure of the head domain of Burkholderia pseudomallei antigen BPSL2063. Acta Crystallogr D Biol Crystallogr. 2015 Nov;71(Pt 11):2227-2235.
32. Holz C, Lueking A, Bovekamp L, Gutjahr C, Bolotina N, et al. A human cDNA expression library in yeast enriched for open reading frames. Genome Res. 2001 Oct;11(10):1730-1735.
33. Anton BP, Raleigh EA. Complete Genome Sequence of NEB 5-alpha, a Derivative of Escherichia coli K-12 DH5α. Genome Announc. 2016 Nov;4(6).pii: e01245-16.